SAHARA: an online service for HAREM Named Entity Recognition Evaluation

Hugo Gonçalo Oliveira[1], **Nuno Cardoso**[2]

hroliv@dei.uc.pt, ncardoso@xldb.di.fc.ul.pt

[1]CISUC, Universidade de Coimbra, Portugal

[2]LaSIGE, Faculdade de Ciências, Universidade de Lisboa, Portugal

Abstract. *This paper presents SAHARA, an online service for the evaluation platform of Second HAREM. SAHARA allows a fast evaluation of any NER system that conforms with HAREM guidelines, making it easier to perform post-hoc evaluations and keep track of the overall performance of NER systems.*

Resumo. *Neste artigo é apresentado o SAHARA, um serviço na rede para a avaliação do Segundo HAREM. O SAHARA permite a avaliação rápida de qualquer sistema de REM que siga as directivas do HAREM.*

1. Introduction

IR evaluation tools and interfaces, such as TREC-Eval tool[1] or the DIRECT [Dussin and Ferro 2008] interface, are quite useful for the overall IR community, allowing post-hoc evaluations and comparison of system performances according to the evaluation environments. In the NLP area, the recent ANNALIST tool [Demetriou et al. 2008] also aims to cover the evaluation of several semantic annotation systems, NER included. ANNALIST facilitates the access to evaluation platforms, encouraging the community to adopt common practices of evaluating their systems. This is in fact a relevant initiative, specially in the NLP area with its multitude of challenges and problems.

Second HAREM [Mota and Santos 2008] is the second edition of the HAREM joint evaluation on Portuguese NER, organised by Linguateca. HAREM addressed the NER evaluation with a new semantic model, with significant differences in comparison to previous initiatives such as MUC [Hirschman 1998] or ACE [Doddington et al. 2004], and therefore required the development of a specific NER evaluation platform. This paper presents SAHARA, an online service that allows fast access to HAREM's evaluation platform, which is publicly available through the URL `http://www.linguateca.pt/HAREM/` → Avaliador.

2. HAREM's evaluation architecture

HAREM's evaluation platform relies on a modular architecture, briefly presented in Figure 1 and described in detail in [Gonçalo Oliveira et al. 2008]. Besides the classical NER track, Second HAREM had two more evaluation tracks, both dependent on NER: TEMPO track [Baptista et al. 2008] (identification of additional attributes for time related NEs) and ReRelEM [Freitas et al. 2009] track (identification of semantic relations between

[1]TREC - `http://trec.nist.gov`

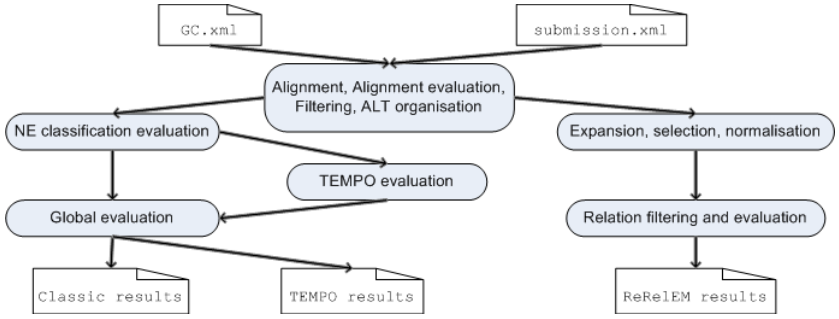

Figure 1. The evaluation platform.

NEs). A golden collection (GC) was used for each track, each consisting of a manually annotated XML file with a subset of the documents in the Second HAREM collection.

In the first stage of the evaluation, similar to all three tracks, the NEs in the GC are aligned with the NEs in the participation and their delimitation is evaluated. Participants could design their own selective scenario, consisting of a subset of categories. SAHARA enables evaluation of each participation in all possible scenarios. Another feature of Second HAREM, taken care in this stage, is the possibility of using special tags (ALT) to identify different alternatives in vague NE delimitation.

In the second stage, evaluation follows different paths, depending on the track. In classic HAREM, NE classification is evaluated by comparing category, type and subtype values of each NE in the GC with the same values in the corresponding submitted NE and applying the measure in [Gonçalo Oliveira et al. 2008]. For the TEMPO track, additional attributes of time related NEs are considered. As for ReRelEM, the relations in both GC and participation are first expanded (according to inversion and transitivity rules), their arguments are normalised, converted into triples and one point is finally given for each correct relation. For the three tracks, global results are computed, including precision, recall and F-measure of the participation.

3. SAHARA

SAHARA is a web interface to the HAREM evaluation platform, that performs post-hoc NER evaluations in three steps: (i) run upload, (ii) configuration, and (iii) result display.

3.1. Uploading a run

SAHARA's homepage starts by prompting the user to upload a submission file in HAREM's XML format (eventually compressed with ZIP), or to insert a file URL.The file is validated with RelaxNG[2] rules on its syntax according to HAREM guidelines. Valid runs are allowed to move on to the evaluation configuration, while runs with errors are not accepted – RelaxNG errors are listed and SAHARA prompts for a new upload.

3.2. Configuring the evaluation

While the classic HAREM track is mandatory, the user can choose to include the TEMPO track and/or the ReRelEM track on the evaluation (see Figure 2). For each selected track,

[2]RelaxNG - `http://relaxng.org/`

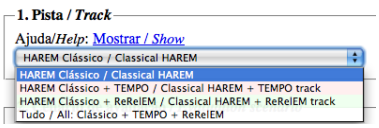

Figure 2. Selection of the evaluation tracks.

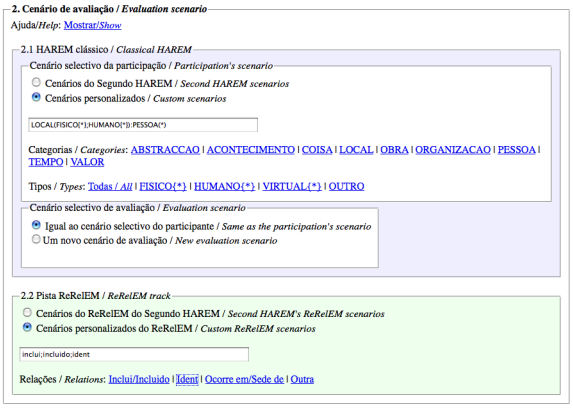

Figure 3. Selection of the evaluation scenario.

the evaluation scenario, the evaluation mode and the GC can be selected. First, the evaluation scenario configures the group of categories to be used in the evaluation (see Figure 3). The user can choose from pre-defined scenarios used in Second HAREM – for example, `LOCAL(FISICO{*}:HUMANO{*})` uses only NE of administrative and physical places –, or he can personalise his evaluation scenario. For ReRelEM, the user can also choose the set of relations, in a similar way. Second, the evaluation mode setups advanced evaluation options (see Figure 4). For the classic track the user can select the ALT tag handling strategy or change the weights on the evaluation measure. In the TEMPO track it is possible to configure the normalisation options for the time entities, while in ReRelEM the expansion of the relations can be controlled to be done only in the GC or also in the participation. Finally, the user can choose the GC to be used on the evaluation (see Figure 5).

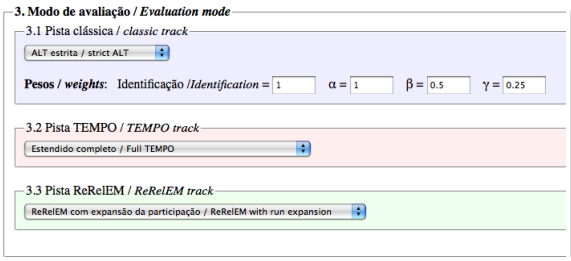

Figure 4. Selection of the evaluation mode.

Figure 5. Selection of the golden collection.

3.3. Displaying the results

According to the chosen evaluation tracks and configuration, SAHARA computes an execution plan where each evaluation module is invoked sequentially. The results are finally presented to the user. For each track, SAHARA summarises the results in a table and uses the precision, recall and f-measure values for a graph comparison with the best three NER participating systems in Second HAREM[3] (see Figure 6). The outputs of each evaluation module can also be inspected (see Figure 7).

[3]Provided that there are official results for the selected evaluation scenario, mode and GC

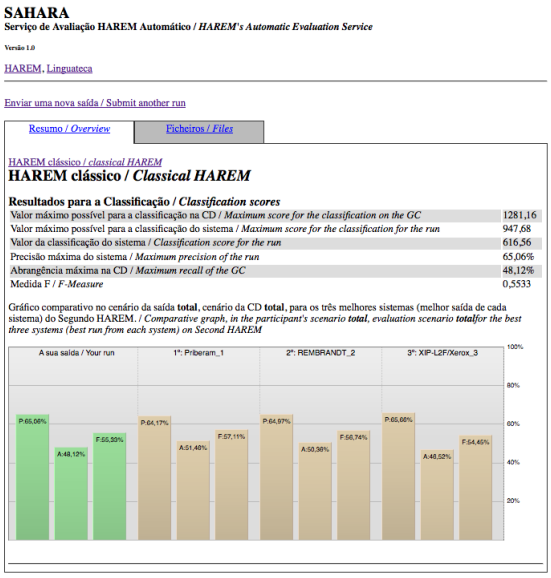

Figure 6. Presentation of the results.

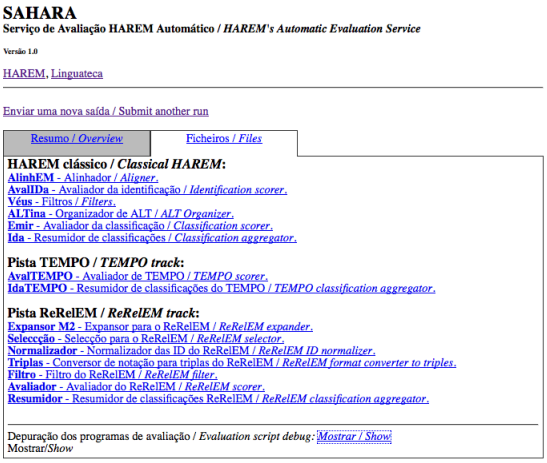

Figure 7. Listing of the output files.

Acknowledgments

SAHARA was developed in the scope of HAREM, a subproject of Linguateca, jointly funded by the Portuguese Government, the European Union (FEDER and FSE), under contract ref. POSC/339/1.3/C/NAC, by UMIC and FCCN. Nuno Cardoso is supported by FCT scholarship grant SFRH/BD/45480/2008. We thank the whole HAREM team.

References

Baptista, J., Hagège, C., and Mamede, N. (2008). Identificação, classificação e normalização de expressões temporais do português: A experiência do Segundo HAREM e o futuro. In *[Mota and Santos 2008]*, pages 33–54.

Demetriou, G., Gaizauskas, R., Sun, H., and Roberts, A. (2008). Annalist - annotation alignment and scoring tool. In *Procs. of LREC'08*, Marrakech, Morocco. ELRA.

Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., and Weischedel, R. (2004). The Automatic Content Extraction (ACE) Program. Tasks, Data and Evaluation. In *Procs. of LREC'04*, pages 837–840.

Dussin, M. and Ferro, N. (2008). Direct: applying the dikw hierarchy to large-scale evaluation campaigns. In *Procs. of JCDL'08*, pages 424–424, New York, NY, USA.

Freitas, C., Santos, D., Mota, C., Gonçalo Oliveira, H., and Carvalho, P. (2009). Detection of relations between named entities: report of a shared task. In *Semantic Evaluations: Recent Achievements and Future Directions (SEW), NAACL-HLT Workshop*.

Gonçalo Oliveira, H., Mota, C., Freitas, C., Santos, D., and Carvalho, P. (2008). Avaliação à medida no Segundo HAREM. In *[Mota and Santos 2008]*, pages 97–129.

Hirschman, L. (1998). The evolution of evaluation: Lessons from the Message Understanding Conferences. *Computer Speech and Language*, 12(4):281–305.

Mota, C. and Santos, D., editors (2008). *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca.